**ISSN 2710-1673. Artificial Intelligence, 2020, № 3**

**UDC 519.8** DOI: https://doi.org/10.15407/jai2020.03.065

***T.O. Bardadym[1], V.M. Gorbachuk[1], N.A. Novoselova[2], C.P. Osypenko[1], V.Yu. Skobtsov[2]***
[1]V.M.Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine, Ukraine
Academician Glushkov Ave., 40, Kyiv, 03187
[2]United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Belarus
6, Surganava St., Minsk, 220012

# INTELLIGENT ANALYTICAL SYSTEM AS A TOOL TO ENSURE THE REPRODUCIBILITY OF BIOMEDICAL CALCULATIONS

***Т.О. Бардадим[1], В.М. Горбачук[1], Н.А. Новоселова[2], С.П. Осипенко[1], В.Ю. Скобцов[2]***
[1]Інститут кібернетики імені В.М. Глушкова НАН України, Україна
пр. Академіка Глушкова, 40, м. Київ, 03187
[2]Об'єднаний інститут проблем інформатики НАН Білорусі, Білорусь
вул. Сурганова, 6, м. Мінськ, 220012

# ІНТЕЛЕКТУАЛЬНА АНАЛІТИЧНА СИСТЕМА ЯК ІНСТРУМЕНТ ЗАБЕЗПЕЧЕННЯ ВІДТВОРЮВАНОСТІ БІОМЕДИЧНИХ ОБЧИСЛЕНЬ

**Abstract**. The experience of the use of applied containerized biomedical software tools in cloud environment is summarized. The reproducibility of scientific computing in relation with modern technologies of scientific calculations is discussed. The main approaches to biomedical data preprocessing and integration in the framework of the intelligent analytical system are described. At the conditions of pandemic, the success of health care system depends significantly on the regular implementation of effective research tools and population monitoring. The earlier the risks of disease can be identified, the more effective process of preventive measures or treatments can be. This publication is about the creation of a prototype for such a tool within the project «Development of methods, algorithms and intelligent analytical system for processing and analysis of heterogeneous clinical and biomedical data to improve the diagnosis of complex diseases» (M/99-2019, M/37-2020 with support of the Ministry of Education and Science of Ukraine), implemented by the V.M. Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine, together with the United Institute of Informatics Problems, National Academy of Sciences of Belarus (F19UKRG-005 with support of the Belarussian Republican Foundation for Fundamental Research). The insurers, entering the market, can insure mostly low risks by facilitating more frequent changes of insurers by consumers (policyholders) and mixing the overall health insurance market. Socio-demographic variables can be risk adjusters. Since age and gender have a relatively small explanatory power, other socio-demographic variables were studied – marital status, retirement status, disability status, educational level, income level. Because insurers have an interest in beneficial diagnoses for their policyholders, they are also interested in the ability to interpret relevant information – upcoding: insurers can encourage their policyholders to consult with doctors more often to select as many diagnoses as possible. Many countries and health care systems use diagnostic information to determine the reimbursement to a service provider, revealing the necessary data. For processing and analysis of these data, software implementations of construction for classifiers, allocation of informative features, processing of heterogeneous medical and biological variables for carrying out scientific research in the field of clinical medicine are developed. The experience of the use of applied containerized biomedical software tools in cloud environment is summarized. The reproducibility of scientific computing in relation with modern technologies of scientific calculations is discussed. Particularly, attention is paid to containerization of biomedical applications (Docker, Singularity containerization technology), this permits to get reproducibility of the conditions in which the calculations took place (invariability of software including software and libraries), technologies of software pipelining of calculations, that allows to organize flow calculations, and technologies for parameterization of software environment, that allows to reproduce, if necessary, an identical computing environment. The main approaches to biomedical data preprocessing and integration in the framework of the intelligent analytical system are described. The experience of using the developed linear classifier, gained during its testing on artificial and real data, allows us to conclude about several advantages provided by the containerized form of the created application: it permits to provide access to real data located in cloud environment; it is possible to perform calculations to solve research problems on cloud resources both with the help of developed tools and with the help of cloud services; such a form of research organization makes numerical experiments reproducible, i.e. any other researcher can compare the results of their developments on specific data that have already been studied by others, in order to verify the conclusions and technical feasibility of new results; there exists a universal opportunity to use the developed tools on technical devices of various classes from a personal computer to powerful cluster.

**Keywords:** classifier; cloud service; containerized application; gene expression data; isolated software environment; reproducibility of calculations; biomarker.

 

**Анотація**. Підсумовано досвід використання прикладних контейнеризованих біомедичних програмних засобів у хмарному середовищі. Вказано шляхи забезпечення відтворюваності наукових обчислень при використанні сучасних технологій наукових розрахунків. Описано основні підходи до попередньої обробки та інтеграції біомедичних даних у рамках інтелектуальної аналітичної системи. В умовах пандемії успіхи системи охорони здоров'я суттєво залежать від регулярного впровадження ефективних засобів досліджень і моніторингу стану населення. Чим раніше вдається виявити ризики появи захворювання, тим ефективніше може йти процес профілактичних заходів або лікування. У даній публікації йдеться про створення прототипу такого засобу в рамках проєкту «Розробка методів, алгоритмів і інтелектуальної аналітичної системи для обробки й аналізу різнорідних клінічних та біомедичних даних з метою вдосконалення діагностики складних захворювань» (М/99-2019, М/37-2020 за підтримки Міністерства освіти та науки України), що виконується Інститутом кібернетики імені В.М.Глушкова НАН України спільно з Об'єднаним інститутом проблем інформатики НАН Білорусі (Ф19УКРГ-005 за підтримки Білоруського республіканського фонду фундаментальних досліджень). Страховики, що входять у ринок, можуть страхувати переважно низькі ризики, сприяючи частішим змінам страховиків з боку страхувальників і змішуючи загальний ринок страхування. Коригувачами ризику можуть бути соціально-демографічні змінні. Оскільки вік і стать мають відносно невелику пояснювальну спроможність, то вивчалися інші соціально-демографічні змінні − сімейний статус, пенсійний статус, статус інвалідності, освітній рівень, рівень доходу. Оскільки страховики мають інтерес до вигідних діагнозів для своїх страхувальників, то також мають інтерес до можливостей трактування відповідної інформації − перекодування інформації: страховики можуть заохочувати своїх страхувальників консультуватися з лікарями, щоб відбирати більше діагнозів. Багато країн і систем охорони здоров'я використовують діагностичну інформацію для визначення відшкодування провайдеру відповідних послуг, відкриваючи необхідні для цього дані. Для обробки й аналізу цих даних розробляються програмні реалізації побудови класифікаторів, виділення інформативних ознак, опрацювання різнорідних медико-біологічних змінних для проведення наукових досліджень у галузі клінічної медицини. У статті підсумовано досвід використання прикладних контейнеризованих біомедичних програмних засобів у хмарному середовищі. Вказано шляхи забезпечення відтворюваності наукових обчислень при використанні сучасних технологій наукових розрахунків. Зокрема, увага привертається до контейнеризації біомедичних додатків (технології Docker, Singularity), за рахунок чого досягається відтворюваність середовища для виконання обчислень (використання ідентичних програмних засобів та бібліотек), технології конвеєризації, що допомагає організувати обчислення в потоковому режимі, та технології параметризації обчислювального середовища, що дозволяє, за необхідності, створювати ідентичне обчислювальне середовище. Описано основні підходи до попередньої обробки та інтеграції біомедичних даних в рамках інтелектуальної аналітичної системи. Досвід використання розробленого лінійного класифікатора, набутий при його тестуванні на штучних та реальних даних, дозволяє зробити висновок про декілька переваг, які надає контейнеризована форма створеного додатку: вдається забезпечити доступ до реальних даних, розташованих у хмарних середовищах; забезпечується можливість виконання обчислень для розв'язування дослідницьких задач на хмарних ресурсах як за допомогою розроблених засобів, так і за допомогою хмарних сервісів; така форма організації досліджень робить числові експерименти відтворюваними, тобто будь-який інший дослідник може порівняти результати роботи своїх розробок на конкретних даних, які вже вивчали інші, з метою перевірити зроблені висновки та технічні можливості нових розробок; з'являється універсальна можливість використовувати розроблені засоби на технічних пристроях різного класу від персонального комп'ютера до потужного кластера.

**Ключові слова:** класифікатор; хмарний сервіс; контейнеризований додаток; дані експресії генів; ізольоване програмне середовище; відтворюваність обчислень; біомаркер.

## Introduction

This publication summarizes the experience of the use of applied containerized software tools in cloud environment, which the authors gained during the project «Development of methods, algorithms and intellectual analytical system for processing and analysis of heterogeneous clinical and biomedical data in order to improve the diagnosis of complex diseases», accomplished by the team from the United Institute of Informatics Problems of the NAS of Belarus and V.M.Glushkov Institute of Cybernetics of the NAS of Ukraine. The main approaches and program tools for the development of intellectual analytical system are described.

## Problem formulation

At the conditions of pandemic, the success of health care system depends significantly on the regular implementation of effective research tools and population monitoring [1−2]. Decisions on which the people's lives depend are regularly made not only by individuals, but also by legislative and executive institutions of power that implement the function of state health care system. These decisions take into account the possibility of preserving and prolonging human life by means of scarce resources (for example, financial, human, temporal resources). Such decisions are made by government institutions to implement the functions of defence and security,

 

law and order, macroeconomic management, protection of property rights.

**Analysis of modern researches and publications**

The countries, having a national health service or national health insurance, usually allow government agencies to make decisions on orders for new products (pharmaceuticals, therapies and medical devices). As a rule, innovations, that promote therapeutic treatment with a lower probability of early death within a certain risk group, predominate [3]. Because such innovations involve additional costs, cost-cutting innovations are often neglected.

For instance, providing a multimillion-dollar mobile coronary unit can help treat patients with heart attacks quickly, significantly reducing the number of lethal cases on the way to a hospital. The long-term drug therapy for patients with hypertension who use antihypertensive drugs can also prevent heart attacks, significantly supporting the economy of research and development (R&D) in pharmaceuticals. The installation of dialysis equipment for patients with chronic renal failure promotes R&D in manufacturing medical equipment.

**Goal of the research**

The earlier the risks of disease can be identified, the more effective process of preventive measures or treatments can be [4]. Life-saving costs are borne not only in the field of health care: in the field of transport, in locations with a higher number of road accidents, there are issues of improving the quality of roads (not only road surface), which must be met by local communities and government agencies; in the field of transport, there are also issues of proper arrangement of roads within residential areas in order to reduce speed of vehicles and to conduct permanent video surveillance. Of course, the practical realization of responses to those issues involves certain expenditures of the local or state budget.

**Main results**

In the field of environmental protection, there are questions about ensuring the levels of security systems for such dangerous enterprises as a nuclear power plant or a chemical plant; if the level of security system is insufficient, an accident can occur threatening the lives of millions of people. One of the consequences of the 1986 Chornobyl disaster was an increase in cancer cases, especially in Ukraine and Belarus. In thermal power plants burning coal, there are questions about the cost of filters that can contain sulfur dioxide and other harmful emissions into the atmosphere. Such emissions increase the incidence of respiratory diseases among the people.

In all the above issues, government institutions cannot make rational decisions without a comprehensive and accurate assessment of future gains (and losses) caused by the implementation of a particular project, as well as without comparison of such gains with the present value of cost flow associated with the project. It is important for decision makers to measure gains and costs in the same units. Since project costs are usually measured in monetary terms, it makes sense to measure all gains in monetary terms as well. Therefore, the prolongation of life or improvement of human health, caused by the implementation of project should also be measured in monetary units. Since it is difficult to assess the status of health and life for a human being in monetary units, economists have developed alternative methods for assessing the state of health and human life.

Different approaches to economic health assessment compare the benefits of medical intervention with the costs of this intervention. Gains from intervention can be measured by physical units on a one-dimensional scale, monetary units, units of cardinal utility function reflecting the multidimensional concept of health in a scalar index.

Since the 1990-s, several states in the world have taken steps to increase competition for their health care insurers, hoping to improve efficiency in their fields of health insurance and health care. Then the generalized equality of price and marginal cost will mean that competing health insurers will charge a high premium for high risks and at the same time a low premium for low risks: high risks

are characterized by a relatively high expected cost of treatment due to the high probability of disease. As the state wants all its citizens to be provided with health insurance, there are issues of risk selection in health insurance markets.

One way to ensure an universal access to health insurance is to provide targeted subsidies to the poorer strata of population to cover insurance premiums. In practice, governments regulate premiums, effectively eliminating the dependence of premium charged by an insurer on risk: in the United States, for example, premium regulation applies so called a community rating. In addition, the German and Swiss regulators typically require insurers to follow an open enrollment policy and accept all the applications. In the United States Medicare gives its beneficiaries a choice between the Medicare Plan itself and competing health care plans, which receive a capitation payment for every policyholder.

Therefore, in the countries mentioned, there is a natural incentive to risk selection. If each person pays the same insurance premium, the insurer will expect losses with high-risk individuals (of high-risk type) and gains with low-risk individuals (of low-risk type). The economic viability and balance of any health insurer presumes a sufficient number of low-risk persons insured: insurers try to attract as many such persons as possible. Therefore, under the pressure of competition, all the insurers will take part in the collection of cream on market (cream-skimming), attracting favorable risks and avoiding adverse risks.

Risk selection can take many forms. On the one hand, health insurers can implement direct risk selection by influencing who would sign the insurance contract: for example, the insurers may not pay their attention to the draft contract from a high-risk person. Individuals who are likely to need some medical care may be asked to sign a contract that provides additional discount services or outright payments. On the other hand, indirect risk selection is the development of payment packages or contracting with service provides that involve low-risk individuals but do not involve high-risk persons. Direct risk selection concerns the problem of individual access to a service, and indirect one − the quality problem.

The both forms of risk selection will occur only when insurers or their consumers possess information about individual health care costs. Direct risk selection require insurers to be able to observe the characteristics of physical persons that correlate with their expected costs − gender, age, social behaviour, and so on. For instance, if healthy people use the Internet more often, the risk selection strategy is to market insurance contracts online: this way people do not have to know their type of risk. However, people need to know their type of risk in indirect risk selection: for example, people need to know the likelihood that they will use certain services. Such personal data allow insurers to develop payment packages and attract service providers with different types of risk.

Direct and indirect risk selection can take place simultaneously: measures that exclude one selection should not affect another. For instance, if the benefit package is strictly regulated, preventing indirect risk selection, insurers may remain interested in attracting favorable risks and thus turn to another risk selection − direct risk selection. On the contrary, if insurers do not have the ability to select risks directly, they retain the incentive to develop a benefit package that attracts low risks and avoids high risks. Indirect risk selection is closely related to the phenomenon of unfavorable (adverse) selection in insurance markets, which happens when policyholders have more information about their type of risk in comparison with their insurers. This phenomenon takes place regardless of the actions of state. At the same time, indirect risk selection is an implication of state regulations for premiums.

To avoid unwanted behavior by insurers in selecting risks, certain measures can be taken based on the assumption of compulsory health insurance, forcing them to cover high risks by means of low risks.

First, open enrollment guarantees that some insurers will take some high risks. At the same time, legislation, regulation and repor-

 

ting may prevent obvious opportunities for direct risk selection: for example, the law may limit the insurer's financial and other benefits from taking low risks.

Second, the measure against indirect risk selection is the regulation of benefit package. On the one hand, lower bounds of benefits can be envisaged, forcing insurers to offer benefits that are important for high risks (say, for the treatment of different types of diabetes). On the other hand, upper bounds of payments may prevent insurers from including low-risk services (say, fitness center services) in their contracts. In addition, certain types of payments, that are convenient for risk selection, can be regulated by separate provisions. However, the payment package includes supply of services from specific partners provided for in the contract (say, subcontractors), which may be selected by the insurer in question. Such selection is especially important in Managed Care: for example, by involving many sports medicine professionals, the insurer can count on the attention of healthy lifestyle advocates (low-risk consumers).

Third, the measure of creating incentives via additional payments to insurers, who take high risks, and imposing financial sanctions to insurers, who skim creams (favorable risks), is a risk adjustment scheme (RAS). The payments mentioned depend on such characteristics observed as age and gender. The measure of reimbursing the share of actual costs for medical treatment is a cost reimbursement scheme (CRS). The idea of CRS is to reduce gains from risk selection by decreasing the impact of costs on the profits of insurers. At the same time, the CRS reduces incentives of insurers to control their costs.

The RAS and CRS can be substantiated by modeling risk selection. First of all, due to various reasons insurers may differ in their terms of insurance for population, the RAS and CRS can create a competitive system where the favorable risk structure of an insurer does not give her a starting advantage. Besides, the health insurance market may be destabilized as new insurers enter the market and move from high to low risks. The RAS and CRS can reduce differences of insurers in premiums, thereby reducing incentives to the movement (transition).

The insurers, entering the market, can insure mostly low risks by facilitating more frequent changes of insurers by consumers (policyholders) and mixing the overall health insurance market. Because insurers, that have entered the market earlier, would appear at high risks, they eventually have to increase their premiums or file for bankruptcy. In such circumstances, insurers will have no incentive to invest in proving effective payments.

Indeed, there is evidence of higher low-risk mobility in the German health insurance market, based on a comparison of the health care expenditure (HCE) of those who change insurers and those who do not change their insurer: depending on age categories, people, who changed insurers, had on average 45−85 % less HCE than the HCE of those who did not change insurers. Studies, based on the German socioeconomic panel, have shown that (adult) people, who remained loyal to their insurer, had significantly worse health status than people who changed insurers. In the United States, there is a case of Harvard University's decision to increase employers' contributions to insurance premiums if employers did not choose the cheapest option (Health Maintenance Organization (HMO) plan).

Types of risk began to be identified during the year: those who switched from the most expensive insurance plans to HMOs had a mean age of 46 years and were 9% higher in HCE than the overall average HCE; those who remained on expensive insurance plans had an average age of 50 years and a 16% higher HCE compared to the general average HCE. The rapid loss of low risks by broad insurance plans forced the experiment to stop.

Thus, the RAS and CRS can help ensure a level playing field during the transition to a competitive market and the stabilization of health insurance market. In the absence of schemes such as the RAS and CRS, the market may lose the most efficient insurers. For actuaries and other financial professionals, risk

adjustment means the accrual of a premium or per capita payment in proportion to the expected expenses of an individual or group. The RAS is based upon risk adjusters – the observed characteristics of individuals. The development of RAS and the search for appropriate risk adjusters require empirical testing of their ability to predict HCE.

Socio-demographic variables can be risk adjusters. Since age and gender have a relatively small explanatory power, other socio-demographic variables were studied – marital status, retirement status, disability status, educational level, income level. Data from the German health insurance funds showed that elderly pensioners with disabilities have significantly higher HCE. In addition, higher HCEs are revealed by single retirees and low-income individuals.

HCE in previous periods is an obvious indicator of morbidity: an increase in HCE leads to an increase in HCE in the next period by 20−30%. At the same time, the explanatory capacity of HCE should be weighed against the weakening of person's incentives to reduce her costs, because higher current HCE will to some extent be compensated to the person later. It is through HCE that insurers try to identify favorable risks, and there may not be better risk adjusters. Prescription medications in previous periods have predicted the value of HCE.

The morbidity can be measured by gathering available diagnostic information to identify chronically ill patients and to classify individuals according to their expected HCE. This classification can be done by various methods. The empirical studies show that diagnostic information gives an accurate predictiton of HCE values. In turn, the corresponding gathering of information can be expensive. Because insurers have an interest in beneficial diagnoses for their policyholders, they are also interested in the ability to interpret relevant information – upcoding: insurers can encourage their policyholders to consult with doctors more often to select as many diagnoses as possible.

Many countries and health care systems use diagnostic information to determine the reimbursement to a service provider, revealing the necessary data. For processing and analysis of these data, software implementations of construction for classifiers, allocation of informative features, processing of heterogeneous medical and biological variables for carrying out scientific research in the field of clinical medicine are developed.

One of the goals of research includes the development of approaches and program tools for the purpose of the reproducibility of numerical experiments, which were conducted in the framework of the joint project. The goal of the project is to develop effective methods and software for constructing classifiers, selection of informative features, creation of a prototype of an intelligent analytical system, which is a software implementation of all stages of data processing and analysis and is aimed at conducting research in the field of clinical medicine. This system will implement the functions of integrating clinical and molecular patient data, determining diagnostic biomarkers and their combinations, building classifiers of complex diseases (oncological diseases) based on integrated data, identifying new disease subtypes to improve treatment methods and increase its efficiency. The second goal includes the development of the approaches to

Large amount of research activities devoted to the development of mathematical methods of data handling, particularly classification models, is due, on the one hand, to a wide range of possible applications, and on the other hand – the complexity of these problems, which requires the development and improvement of means to solve them (see, for example, [5−9]). In addition to general requirements for efficiency of the created software there exists a need to pay attention to the conditions of availability of large and heterogeneous data sets, requirements for the ability to transfer programs from one hardware to another, their performance in cloud computing.

 

Moreover, one of the most important requirements is the reproducibility of research numerical experiments. The principle of reproducibility of research is one of the basic scientific principles. However, a crisis called "reproducibility crisis" has been realized in science [10−11]. This crisis has affected almost all branches of science, in particular, to a large extent – biology and medicine. Much effort has been made recently to overcome this crisis, including the development of software and software platforms to ensure the reproducibility of scientific computing. Computing in biology and medicine involves the use of high-performance computing technologies (including clusters and grid technologies). However, the introduction of modern technologies to ensure the reproducibility of calculations in this area is quite slow [12, p. 731]. As a result, in the field of cluster technologies, which do not have the appropriate software installed, there is a contradiction between modern requirements for the reproducibility of scientific calculations and the ability to achieve it by old means.

It so happened that the need to create a containerized application was not a planned stage of our study. This was primarily due to the ways of accessing the real data on which the software was tested. Only then did the authors realize that they had gained other advantages, among which the most important is the reproducibility of research numerical experiments. It is the purpose of the publication to share this experience.

The second purpose is to shortly describe our efforts taken towards the development of specialized computer methods and models in order to solve the vital tasks in the field of biomedicine. Nowadays there exists the enormous amount of biomedical and clinical data collected in the public and private repositories. They can be freely accessed and present the wide field for experiments with the newly developed scientific approaches and their comparison. The integration of heterogeneous information sources is one of the urgent applied problems, which we have tried to solve in our project. The hybrid classification model presents the basis of the intelligent analytical system and aims to integrate several sources of biomedical information in order to improve the diagnostics and prognosis of complex diseases.

Based on the approaches presented in [13−14], optimization models and methods for solving problems of constructing linear classifiers have been developed. In particular, the problem of constructing classifiers for linearly indivisible sets was formulated as a problem of minimizing the band of incorrect classification of training sample points. This model belongs to the class of optimization problems of non-convex programming and is multi-extreme. Various formulations of this problem are offered, approaches to construction of approximate decisions and calculation of estimations of optimum values are considered. An interesting geometric interpretation of the problems of constructing linear classifiers can be found in [15].

To solve these optimization problems, methods of non-smooth optimization, namely *r*-algorithms of N.Z. Shor [16−17] and exact penalty functions [18–19] were used. When creating appropriate software, modern libraries of linear algebra, similar to [20−22] should be used to speed up arithmetic operations. It is a combination of algorithms based on non-smooth optimization methods and the use of modern libraries of linear algebra was implemented in the developed software module NonSmoothSVC.

To test the abilities of the new classifier NonSmoothSVC a comparison with existing tools was made. The methods integrated into the library scikit-learn [12; 23] were chosen, namely Linear SVC, NuSVC, Ada Boost. The two last methods are non-linear classifiers; they were chosen to get additional information concerning advantages of different methods for different problems. First numerical experiments were made on specially generated artificial data.

Computational experiments aimed to establish the speed and predictive properties of new software compared to existing ones. Both artificially created data and real medical data were used in the calculations in the test problems. Training and control samples of ran-

domly generated problems were formed as identically distributed data points on a single cube in the space of features $R^n$. Then, the points of the first class shifted in the first coordinate by the value δ, and the points of the second class shifted in the first coordinate by the value (-1-δ). When δ>0, training and control samples are linearly separable, and when δ<0, they are linearly inseparable. Next, the rotation (linear transformation) of space was performed so that the separating hyperplane depended on many coordinates of space. The need to test new software on real data forced us to locate the software module NonSmoothSVC into a containerized application (using Docker technology [24]) for use on a personal computer, as well as on a cluster, grid, and cloud environment.

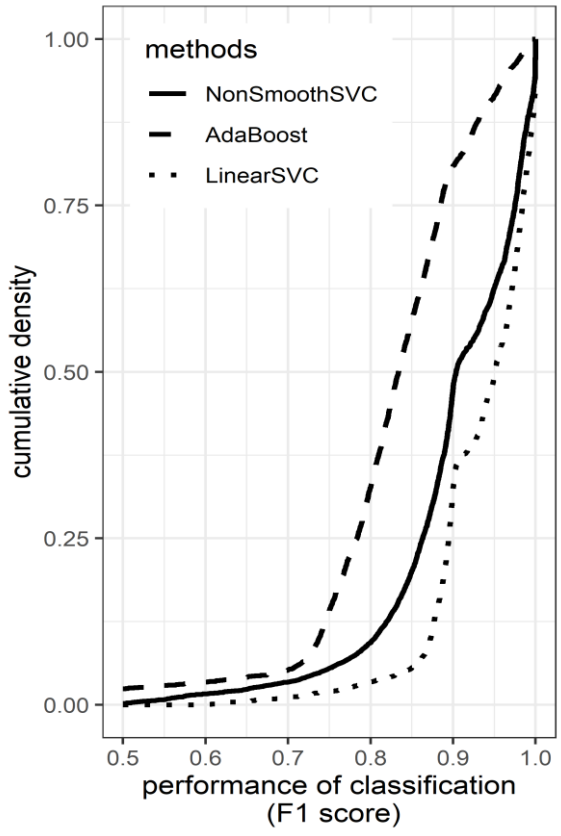


Fig. 1. Comparative density distribution for full data set (*n*=12000)

This permitted to get access to the real data on Cancer Genomics Cloud [25], a specialized cloud platform that provides free access to genetic, medical databases, in particular – The Cancer Genome Atlas (TCGA) [26], and more than 450 public applications designed to analyze data on this topic. It is possible to expand this list with the own applications, data sets, research results (currently there are more than one million on this service), to involve other researchers in projects. Computational experiments have demonstrated that on some data sets the NonSmoothSVC has qualitative advantages over other methods involved in the comparison, but is inferior in speed. Particularly, on linearly separable samples the NonSmoothSVC gained an advantage over the LinearSVC in the number of cases with better classification accuracy.

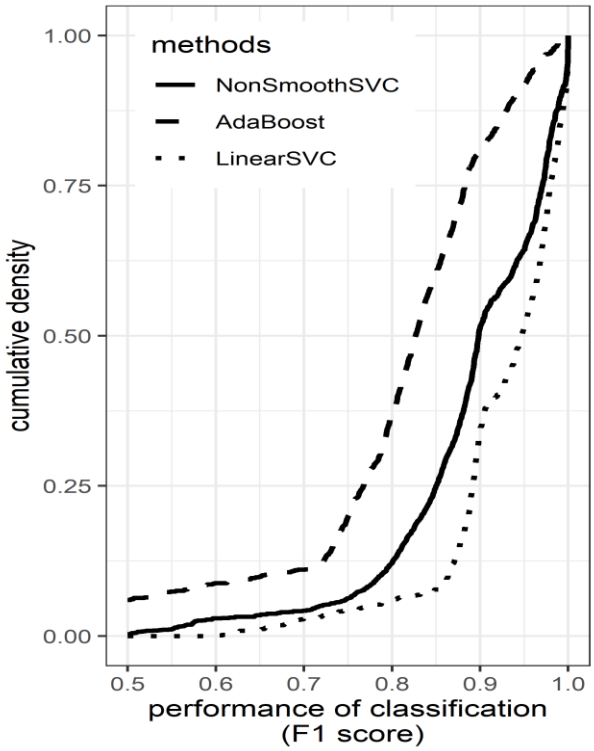


Fig. 2. Comparative density distribution for imbalanced data set (*n*=3720)

On the unbalanced samples, the NonSmoothSVC software slightly outperformed the LinearSVC software in the number of cases with better classification accuracy on average, but demonstrated an advantage in some parts of the classification accuracy scale (Fig. 1–6).

Full description of numerical experiments and the results of testing can be found in the reports (in Ukrainian) at http://moderninform.icybcluster.org.ua/ais/.

 

Thanks to the containerized form, the developed software can become publicly available tools and applications of this and other services in the problems of constructing optimized linear classifiers using modern libraries of linear algebra.

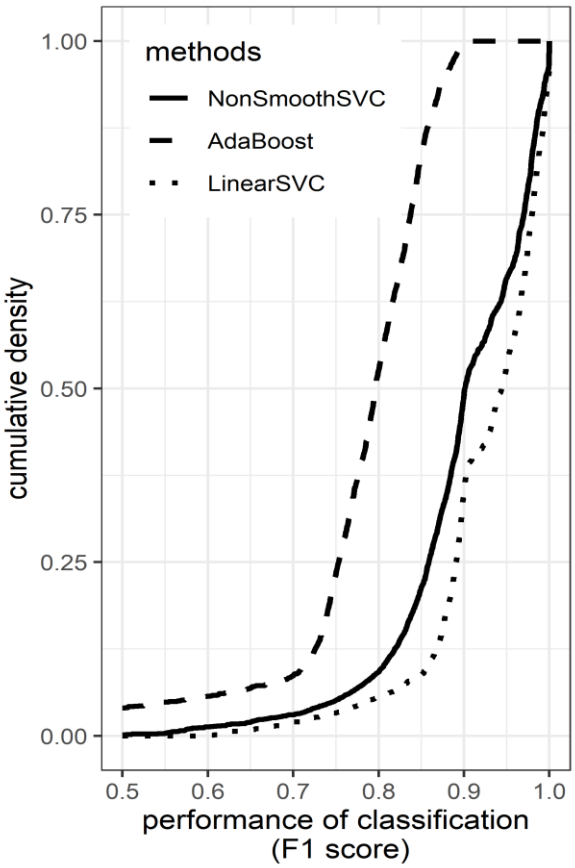


Fig. 3. Comparative density distribution for large data set (*n*=7200)

In the presence of technical possibilities, parallelization on microprocessor networks looks promising.

This approach is especially recommended in the case of large data samples, when the dimension of the feature space is tens of thousands. It is also necessary to take into account the features of optimization problems in specific cases. In particular, additional requirements that may be formulated by specialists may reduce the number of informative features.

Processing and study of biomedical data have some peculiarities. This, in particular, the existence of possible large errors that arise in the processing of medical information and huge number of features that need to be taken into account, which increases the dimensionality of the corresponding optimization problems, the missed measurements, which requires the use of specialized methods for their processing and analysis.

In order to improve the diagnosis and treatment of complex diseases, much attention is paid to the comprehensive analysis of various biomedical and clinical data to understand the processes occurring in the body at the cellular level and changes caused by the development of the disease.

It is known, the cause of complex diseases, along with external factors, is a combination of genetic failures, which does not allow to fix only one genetic mutation as a biomarker. The difficulty also lies in the fact that individual genetic factors can differ and individual cases of the same disease (phenotype) can be caused by different genetic changes. In addition, in the case of the combined effect of several mutations, the individual effect of each of them can be rather insignificant and, therefore, difficult to be detected.

It is also necessary to take into account the high heterogeneity of the complex disease, i.e. heterogeneity of its observed manifestations (phenotypes).

Recently, the methods of systems biology have become widely used to study complex diseases, namely, knowledge about the interactions between genes, their products and small molecules that form a complex network of interactions. This approach makes it possible to explain the appearance of similar phenotypes despite different genetic causes, namely, their interconnection and influence (dysregulation) on the same component of the cellular system. Thus, the use of interactome in conjunction with other data from biogenetic studies can contribute to understanding the processes occurring at the molecular level in complex diseases. The use of combinations of heterogeneous data makes it possible to determine dysregulated cellular pathways, to reveal the relationship between genotype and phenotype, and to explain the heterogeneity of a complex disease.

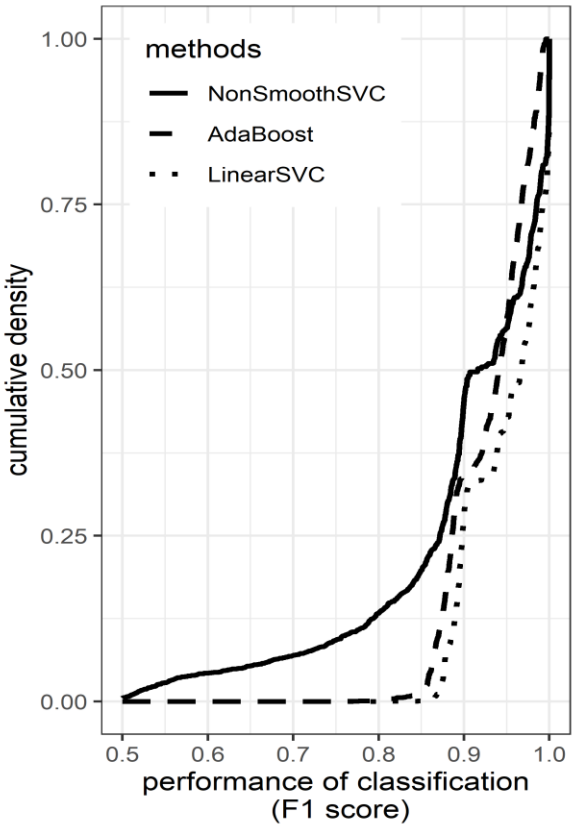


Fig. 4. Comparative density distribution for small data set (*n*=2400)

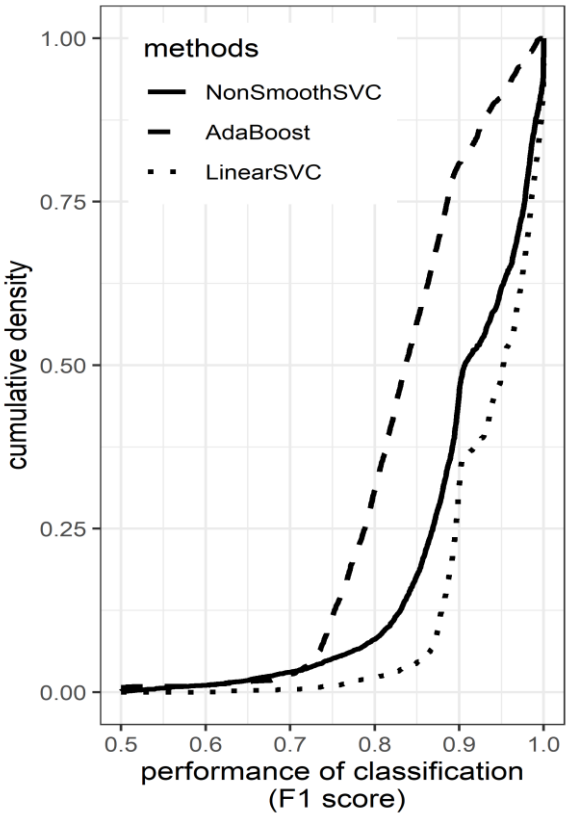


Fig. 5. Comparative density distribution for balanced data set (*n*=8280)

Natural approaches here are: to increase the efficiency of tools to solve such optimization problems and the use of methods for selection informative features. In the works [27−30] attention is paid to the preliminary preparation of available medical data in order to select informative features.

In the course of the project, algorithms for preprocessing and extracting biomarkers from biomedical data were developed, including: an algorithm for ranking features by information content for classification [23]; an algorithm for identifying combinations of biomarkers, taking into account the correlation of features and allowing to exclude their influence.

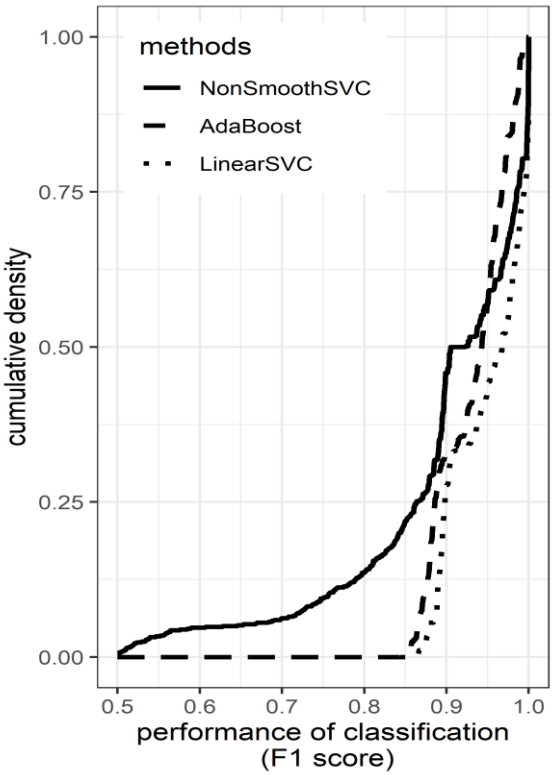


Fig. 6. Comparative density distribution for imbalanced+small data set (*n*=720)

 

Moreover, several approaches were analyzed for identifying a subset of informative features, taking into account several data sources, namely, gene expression data and data on functional and physical interactions of genes and their products, presented in the form of networks. Based on the analysis of existing approaches, an algorithm for identifying a subset of features has been developed, which allows integrating interactomic and transcriptomic data to determine functional subnets associated with the disease. Pre-processing of biomedical data made it possible to reduce the feature space and thereby increase the accuracy of classification models.

Detailed description of algorithms and related information can be found in the report at http://moderninform.icybcluster.org.ua/ais/ (in Russian).

In one of the numerical experiments the real data contained information on the gene expression of cancer patients (143 observations of 60,483 features) obtained from the Cancer Genome Atlas (TCGA). From these data by means of the simplified method of ranking of features proposed by Novoselova [28] 23 most informative features concerning the forecast of a vital status of patients having diagnosed glioblastoma were identified. This approach substantially simplifies numerical difficulties in following data processing.

Due to the fact that various sources of biological information characterize various changes occurring in the body at the cellular level during the development of a complex disease, it is assumed that their combination will improve the accuracy of diagnosis of the subtype of the disease, the reliability of the disease prognosis and response to therapy [31]. In addition, combining heterogeneous data will allow one to discover the relationships between various biomedical entities (genes, proteins, metabolites, etc.) directly related to the development of the disease, compensate for noise and errors in individual data sources and thereby obtain more reliable results. A common problem in solving this problem is how to combine information from different data sources. In our study, of interest are methods for constructing classifiers based on various sources of multidimensional data, which, as a rule, have a heterogeneous representation. Consequently, the task is to unify this representation, determine the base classifier, build classification models on each data source, and select ways to combine the predicted values, obtained using the constructed models.

The core of the intelligent analytical system being developed is a hybrid classification model, which allows combining several sources of biological information about patients in order to build a classification model that allows diagnosing subtypes of complex diseases characterized by genetic disorders. The proposed hybrid model is a classification ensemble with the following distinctive features:

1. Uniform presentation of information from various data sources by constructing a matrix of object-object distances using various kernel functions (density functions), including Gaussian, polynomial function, scalar product of vectors, etc.
2. Implementation of the procedure for selecting classification characteristics for each individual data source.
3. Construction of a basic or individual classifier of a hybrid model, which can be either a single classifier or an ensemble of classifiers built on a single data source.
4. Implementation of several ways of integrating individual classifiers of the model.
5. Analysis of the information content of individual classifiers using the assessment of their weight coefficients.

The method for constructing a hybrid model is based on a combination of the bagging procedure and the aggregation of ranked lists to build basic classifiers and a pruning procedure to determine the final structure of the model, which allows adaptively adjusting the ensemble taking into account the type of classified data.

The preliminary experiments on the TCGA data [26] showed that the ensembles built on heterogeneous data sources can suffi-

ciently increase the accuracy of classification and prediction of subtypes of complex diseases, since each of the data sources describes the organism under study in different planes: gene expression data, Ribonucleic acid (RNA) sequencing, metabolic data, gene copy number data, etc.

Ensuring the reproducibility of calculations is a prerequisite for the reproducibility of scientific research as a whole. The conditions for computational reproducibility are the availability of source data, the ability to reproduce an identical computing environment (or an environment that does not lead to other calculation results), and the availability of the results of computations. Biomedical calculations have their own specific features that should be taken into account when planning them. Let we mention some of them.

Modern biomedical calculations, especially based on genome data, are very huge and cumbersome. Usually "classic" biomedical applications (PAML, Muscle, MAFFT, MrBayes, BLAST, etc.) and large libraries with implementations of biomedical algorithms written in different programming languages (C / C ++, Java, R, Go, Scala, Haskell, Perl, Python, Ruby, Erlang, Julia, etc. [32]) are quite often used simultaneously in one study. Moreover, biomedical calculations often involve methods of artificial intelligence – machine learning, pattern recognition, and corresponding libraries (e.g., scikit-learn [6], [17]). Such a variety of software requires careful configuration of the computing environment with control of the versions of libraries used (here can be used as dozens and hundreds of libraries).

Otherwise one can get a lack of reproducibility as a result of calculations. In terms of using cluster technologies, creating such environments (separate for each user) and maintaining them in a conflict-free state is quite a burdensome task (unless you use special software configuration tools, such as Conda, Bioconda, or containerization of applications using, for example, technology Singularity). Most of the libraries and applications used in biomedical computing do not provide efficient use of parallel multithreaded computing with multicore processors, and at the same time many of them can be applied to an "embarrassingly parallel" model − a model in which individual pieces of data are calculated in parallel by identical instances of computational processes without transferring messages between them (for example, using Apache Hadoop technology) [12].

Taking into account the peculiarities of biomedical computing, reproducibility and their horizontal scaling (the ability to increase the number of identical computing units to solve one problem) can be achieved through the use of containerized applications, software pipeline computing and parameterization of software environment.

*Technologies of containerization of software applications.* Due to the containerization of biomedical applications (Docker, Singularity containerization technology) the following can be achieved: reproducibility of the conditions in which the calculations took place (invariability of software including software and libraries), the possibility of horizontal scaling provided the use of "stunning" model of parallelism in cluster (Singularity) and cloud (using Docker) calculations.

*Technologies of software pipelining of calculations.* Software pipeline allows you to organize flow calculations (calculations in which the inputs and outputs of processes are interconnected). Thanks to the use of tools for automation of flow calculations (workflow engine) such as CWL (Common Workflow Language), GWL (Guix Workflow Language), Snakemake, Nextflow, it is possible to present a specific calculation in the form of a task (text file, as usual, in YAML format or JSON), the results of which can be reproduced [7]. In addition, there are tools that allow you to create / display such tasks in the form of a graph of processes and data flows. An example of such a tool is RABIX (Reproducible Analyzes for Bioinformatics) − a graphical editor for CWL. Some pipeline tools also use containerization (for example, CWL) −

such tasks can be performed both on a personal computer and in a cloud environment. An important feature of streaming automation tools is that the task description syntax allows you to specify the scale of the calculations, indicating the number of resources required. Seven Bridges' product, Cancer Genomics Cloud (CGC, see http://www.cancergenomicscloud.org/), is an example of a cloud software platform for performing reproducible biomedical computations using containerization and pipelining. It is the use of containerization in the creation of an application for the construction of a linear classifier at the V.M. Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine made it possible to conduct testing on real very voluminous medical data located at the CGC.

*Technologies for parameterization of software environment.* Parameterization of the software environment allows you to reproduce, if necessary, an identical computing environment. GNU Guix, Conda, Bioconda are examples of tools that allow you to create an isolated software environment for individual users in a cluster [12].

At present, there exists a range of technologies to ensure the reproducibility of scientific calculations in cloud and cluster environments. This makes it possible to create biomedical applications adapted to these environments. In the result we get computational basis that satisfies modern requirements for computational reproducibility.

The experience of using the developed linear classifier, gained during its testing on artificial and real data, allows us to conclude about several advantages provided by the containerized form of the created application: it permits to provide access to real data located in cloud environment; it is possible to perform calculations to solve research problems on cloud resources both with the help of developed tools and with the help of cloud services; such a form of research organization makes numerical experiments reproducible, i.e. any other researcher can compare the results of their developments on specific data that have already been studied by others, in order to verify the conclusions and technical feasibility of new results; there exists an universal opportunity to use the developed tools on technical devices of various classes from a personal computer to powerful cluster.

## Conclusions

The next steps of the project include development of the common software interface of the experimental prototype of the intelligent analytical system in order to integrate the developed methods and software modules of biomedical data preprocessing, data clustering and classification. It will allow performing all the steps of data analysis from the single framework and conducting research in the field of biomedicine. The hybrid classification model as a core of the intelligent system will make it possible to integrate multidimensional, heterogeneous biomedical data with the aim to better understand the molecular courses of disease origin and development, to improve the identification of disease subtypes and disease prognosis. Much attention will be paid to the experimentation with different computation approaches on real datasets taking into account the reproducibility of results.

## References

1. Knopov P.S., Norkin V.I., Atoyev K.L., Gorbachuk V.M., Kyryliuk V.S., Bila H.D., Samosyonok O.S., Bogdanov O.V. (2020). *Some approaches to the use of stochastic models of epidemiology to the COVID-19 problem.* Kyiv: V.M.Glushkov Institute of Cybernetics, Retrieved from http://incyb.kiev.ua/archives/3988/dejaki-pidhodi-vikoristannja-stohastichnih-modelej-epidemiologii-do-problemi-covid-19/ (In Ukrainian).
2. Gorbachuk V., Gavrilenko S. (2020). *Analysis for dynamics of COVID-19 spreading in Ukraine and neighboring countries on May 1–10, 2020.* Global and regional problems of informatization in society and nature using 2020. Kyiv: National University of Life and Environmental Sciences of Ukraine, 56–60. (In Ukrainian).
3. Gorbachuk V.M., Dunaievskyi M.S., Suleimanov S.-B. (2020). *Management and administration in the field of health care services. Management and administration in the field of services: selected examples.* T.Pokusa, T.Nestorenko (eds.) Opole: Academy of Management and Administration, 268−279. (In Ukrainian).

4. Gorbachuk V.M., Suleimanov S.-B., Batih L.O. (2020). *Decision making criteria in the branch of health care.* Measurement and control in complex systems. Vinnytsia: VNTU, 149–151. (In Ukrainian).
5. Vorontsov K.V. *Mathematical methods of learning by precedents (Machine Learning Theory)* (in Russian), Retrieved from: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf
6. Gupal A.M., Sergienko I.V. *Symmetry in DNA. Methods for Discrete Sequences Recognition.* Kyiv. Naukova Dumka (in Russian).
7. Baldi P., Hatfield W.G. (2011). *DNA Microarrays and Gene Expression. From Experiments to Data Analysis and Modeling.* Cambridge University Press.
8. Kuhn M., Johnson K. (2013). *Applied predictive modeling*. New York: Springer.
9. Heath L.S., Ramakrishnan N. (2010). *Problem solving handbook in computational biology and bioinformatics.* NY: Springer Science & Business Media.
10. Ioannidis J. (2005). *Why Most Published Research Findings Are False*. PLoS Medicine, vol. 2, no. 8, p. 124.
11. Baker M. (2016). *Reproducibility crisis?* Nature, vol. 26, no. 533, 353-66.
12. Strozzi F. et al. (2019). *Scalable workflows and reproducible data analysis for genomics.* Evolutionary Genomics, 2nd ed., New York, NY: Humana Press, 723-745.
13. Zhuravlev Y., Laptin Y., Vinogradov A., Zhurbenko N., Lykhovyd O., Berezovskyi O. (2017). *Linear classifiers and selection of informative features.* Pattern Recogn. and Image Anal., vol. 27, no. 3, 426-432.
14. Zhuravlev Y., Laptin Y., Vinogradov A. (2014). *Comparison of Some Approaches to Classification Problems, and Possibilities to Construct Optimal Solutions Efficiently.* Pattern Recogn. and Image Anal., vol. 24, no. 2, 189-195.
15. Zhurbenko N.G. (2020). *Linear classifier and projection on polytop.* Cybern. Syst. Anal., vol. 56, no. 3, 1-8.
16. Shor N.Z., Zhurbenko N.G. (1971). *A minimization method using the operation of extension of the space in the direction of the difference of two successive gradients.* Cybernetics, vol. 7, 450-459.
17. Shor N.Z. (1998). *Nondifferentiable Optimization and Polynomial Problems.* London: Kluwer Acad. Publ.
18. Laptin Yu.P. (2016). *Exact penalty functions and convex extensions of functions in decomposition schemes in variables.* Cybernetics and Systems Analysis, vol. 52, 85–95. DOI: 10.1007/s10559-016-9803-8.
19. Laptin Yu.P., Bardadym T.A. (2019). *Problems related to estimating the coefficients of exact penalty functions.* Cybernetics and Systems Analysis, vol. 55, no. 3, 400-412. DOI: 10.1007/s10559-019-00147-2.
20. Chang, Chih-Chung; Lin, Chih-Jen LIBSVM – A Library for Support Vector Machines. Retrieved from https://www.csie.ntu.edu.tw/~cjlin/libsvm/.
21. BLAS (Basic Linear Algebra Subprograms). Retrieved from http://www.netlib.org/blas/.
22. LAPACK—Linear Algebra PACKage. Retrieved from http://www.netlib.org/lapack/.
23. Free software machine learning library for the Python programming language. Retrieved from https://scikit-learn.org/stable/index.html
24. Tools for creation of isolated Linux-containers. Retrieved from https://www.docker.com/
25. The Cancer Genomics Cloud. Retrieved from http://www.cancergenomicscloud.org/
26. The Cancer Genome Atlas (TCGA). Retrieved from https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
27. Novoselova N.A., Tom I.E. (2018). *Integrated network approach to protein function prediction.* The Scientific Journal of Riga Technical University. Information Technology and Management Science, vol. 21, 98–103. DOI: 10.7250/itms-2018-0016
28. Tom I.E. (2016). *Information technologies in the analysis of medical data.* Science and innovations, no. 3, 28-31.
29. Novoselova N.A., Tom I.E. (2016). *Method for constructing clusters in genetic data.* Informatika, no.1(49), 64-74.
30. Novoselova N.A., Tom I.E. (2013). *Algorithm for ranking features for detecting biomarkers in gene expression data.* Artificial Intelligence, no. 3, 58-68.
31. Novoselova N.A., Tom I.E., Ablameyko S.V. (2011). *Evolutionary design of the classifier ensemble.* Artificial Intelligence, no. 3, 429-438.
32. Bonnal R. et al. (2019). *Sharing Programming Resources Between Bio* Projects*. Evolutionary Genomics, 2nd ed., New York, NY: Humana Press, 747-766. DOI: 10.1007/978-1-4939-9074-0_25


*Received 10.06.2020*
*Accepted 12.08.2020*